VTQA：通过实体对齐和跨-媒体推理实现视觉文本问答

Kang Chen哈尔滨工业大学 chenkangcs@stu.hit.edu.cn

Xiangqian Wu哈尔滨工业大学 xqwu@hit.edu.cn



图 1：我们数据集中问题-答案对及其对应图像和文本的示例。文本和图像中同一对象的不同表示以相同颜色标示。**例如，文本中的“Elena”**以及图像中对应于**“Elena”**的目标边界框均以红色标出。

摘要

理想的视觉问答需要在视觉与语言的联合空间中进行理解、定位和推理，并可作为场景理解这一人工智能任务的替代指标。然而，大多数现有视觉问答基准仅限于从预定义选项中挑选答案，且对文本关注不足。我们提出一项新挑战，并构建了一个包含 23,781 个问题、基于 10124 对图像和文本的数据集。具体而言，该任务要求模型对同一实体的多媒体表示进行对齐，以实现图像与文本之间的多跳推理，最终使用自然语言回答问题。此挑战旨在开发并评测能够进行多媒体实体对齐、多步推理和开放式答案生成的模型。

1 引言

人工智能系统的长期目标之一，是像人类一样理解复杂的现实世界，而问答（QA）是一项评估人工智能系统理解能力的良好任务。为了回答问题，人们需要从多种模态中提取信息，例如文本、图像，以及知识库、图和表格等结构化数据。此外，人们还需要对齐这些信息，并在不同模态之间进行多步推理。

尽管视觉问答（VQA） [1] 作为一项多媒体问答任务已得到广泛研究，但视觉问答模型仅在回答问题时从图像中提取信息，主要关注场景识别、计数、颜色以及其他
视觉检测任务，这类任务不需要太多逻辑推理，也不涉及不同模态之间的信息分配。直到最近，才有一些尝试将更多模态信息引入视觉问答任务。例如：（1）FVQA [2] 和 OK-VQA [3] 将知识库（KB）与视觉问答任务相结合，这要求具备知识理解和多步推理能力。然而，在现实世界中构建全面的知识库并不容易，因此这类方法难以回答开放式问题；（2）TextbookQA [4] 和 ScienceQA [5] 以教科书为数据来源，涉及文本、图像、表格及其他多模态信息。但这些数据集中的大多数图像都是人工绘制的示意图，而且问题都采用多项选择形式，与现实世界相去甚远。

为解决这些问题，近期出现了一些新数据集，例如 MultiModalQA [6] 和 MuMuQA [7],，它们涉及跨文本、图像等内容进行推理。然而，这些数据集都采用抽取式问答。对于 MultiModalQA 数据集，每张图像都关联着一个维基百科实体，因此图像推理实际上退化为对图像进行排序，并使用排名靠前图像所对应的实体，这削弱了对跨媒体推理能力的需求。至于 MuMuQA，尽管它要求在图像和文本之间进行定位，并进行多跳推理，但仍存在一些问题：(1) 数据全部来自新闻，导致大多数问题都与人类有关；(2) 问题都遵循一种特定模式：首先进行图像实体

定位，然后在新闻正文中寻找答案；(3) 只有 1384 个人工整理的示例，训练数据则由自动生成，导致质量较差且难以训练。当前的多媒体基准数据集仍与真实世界的问答场景相去甚远，也无法很好地衡量人工智能系统的多媒体理解能力。

为弥合这一差距，我们提出了一项名为“视觉文本问答”（VTQA）的全新挑战及其配套数据集。该数据集仅包含需要同时通过图像和文本进行多跳推理的问题。VTQA数据集中的所有标注首先以中文完成，我们也将提供相应的英语翻译。在本项挑战中，我们将分别设置英语和中文赛道，并分别进行排名。

为了回答 VTQA 问题，所提出的模型需要：（1）学习识别图像和文本中与问题相关的实体；（2）对同一实体的多媒体表示进行对齐；（3）在文本与图像之间进行多步推理，并输出开放式答案。VTQA数据集由 10,124 个图像-文本对和 23,781 个问题组成。这些图像来自 MSCOCO 数据集 [8], ，包含各种实体。标注人员需要先根据图像标注相关文本，然后基于图像-文本对提出问题，最后以开放式形式回答问题。我们通过限制文本的最小长度（超过 100 个中文字符），并删除仅凭图像或文本即可回答的问题，来确保文本信息的丰富性和问题的复杂性。

在这项挑战中，模型需要根据给定的图文对回答问题。信息的多样性、多媒体多步推理以及开放式答案，使我们的任务比现有数据集更具挑战性。本挑战旨在开发并评测具备多媒体实体对齐、多步推理和开放式答案生成能力的模型。

2 挑战任务

如图 1 所示，给定一对图像和文本以及一个问题，系统需要用自然语言回答该问题。重要的是，系统需要：(1) 分析问题并找出关键实体；(2) 对齐图像和文本中的关键实体；(3) 根据问题和已对齐的实体生成答案。例如，在图 1 中，问题 Q1 的关键实体是“Elena”。根据文本“金色头发”，我们可以确定图像中从右数第二个人是“Elena”。最后，我们进一步根据图像信息回答“西装”。至于 Q2，这是一个更复杂的问题，需要多次重复前面的步骤才能作答。

3 数据集

3.1 数据集构建

为了推动这项开放式多媒体多跳 VTQA 挑战的发展，我们收集了一个新的数据集。该数据集包含 10124 对图像和文本，以及 23,781 个问题。我们通过新开发的标注界面收集数据。界面中展示了 MSCOCO 数据集中的图像，首先要求标注人员撰写一段与图像相关的文本。随后，标注人员根据图像和文本对提出问题；标注过程要求这些问题不能仅凭图像或文本作答。最后，由其他标注人员标注问题的答案，并指出答案类别。我们将答案分为三类：(1)“YN”表示是非答案；(2)“E”表示答案从文本中提取；(3)“G”表示答案由图像和文本对生成。对于是非答案，标注人员需要用英语标注“yes”或“no”，因为中文中表达“是”或“否”的词语太多，例如，“可以”和“是的”都表示“是”。

3.2 数据集划分

我们将数据集随机划分为训练集、验证集、开发测试集和测试集，每张图像只会出现在其中一个划分中。各划分分别包含 11312、1245、2189 和 9035 个样本。训练集、验证集、开发测试集和测试集以 JavaScript 对象表示法（JSON）格式的文本文件提供，文件名为 train.json、val.json 和 test.json，其中每个数据样本都存储为一个字典。JSON 文件中的属性如下。

- 问题唯一标识符：每个问题的唯一 ID。• 图像本地路径：

3.3 数据集分析

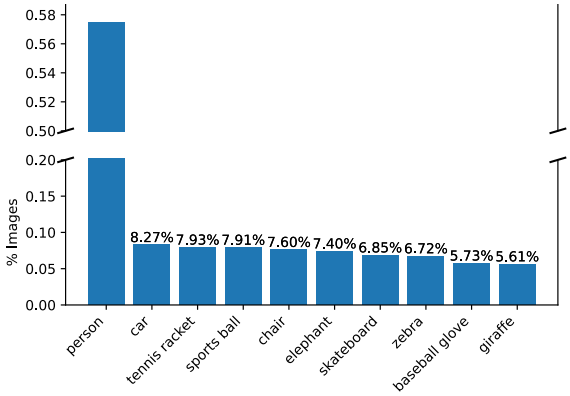

图 2： 我们数据集中所使用图像的前 10 类类别分布。

该数据集中的图像来自 MSCOCO 数据集，其中包含多个物体和丰富的上下文信息。图 2

展示了该数据集中所用图像的前 10 个类别的分布。与 MuMuQA 不同，MuMuQA 中的大多数图像都与人物相关，而我们的数据集中的图像则包含更多种类的物体。

文本和问题的长度统计如图 3 所示。可以看出，大多数文本的长度在 300 到 600 之间，大多数问题的长度在 9 到 21 之间。显然，与以往的视觉问答数据集（通常少于 20 个词）相比，我们的数据集也对文本理解提出了更高要求，以便从长文本中提取信息。

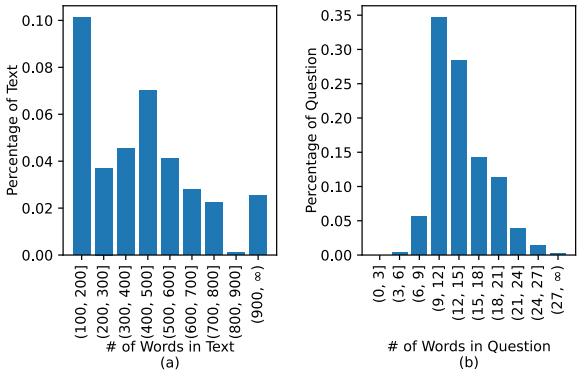

图 3： 不同词长的问题和文本所占百分比。

4 评估标准

由于答案分为三种类型，我们针对不同类型的答案采用不同的指标。

精确匹配（EM）。此指标用于衡量与标准答案完全匹配的预测所占的比例，适用于所有类型的答案。

（宏平均）F1 分数（F1）。此指标用于衡量预测与标准答案之间的平均重叠程度。我们将预测和标准答案视为词元集合，并计算它们的 F1 分数。此指标适用于"E"和"G"类型的答案。

YN准确率（YNAcc）。此指标仅适用于"YN"类型的答案。我们会使用预先定义的是非词典，将答案转换为"是"或"否"。随后，仅根据是非答案计算准确率。

精确匹配指标用于对模型进行排序；其他指标则用于展示算法的性能，不用于排序。

所有训练集和验证数据都将在线发布，供参赛者训练其模型。参赛者有两种数据划分可用于评估模型：（1）测试开发集，该数据集将发布，但不提供对应的答案。参赛者可以预测测试开发集的答案，生成 JSON文件并提交，以快速评估自己的模型。（2）测试集，该数据集将不公开，本次挑战赛的最终排名将依据测试集的总体精确匹配（EM）得分确定。要在测试集上评估模型，参赛者需要将 Docker镜像作为提交内容。我们将提供基线实现，以尽量降低参赛者的额外工作量。我们将在测试集划分上评估模型，并根据总体 EM 指标对参赛者模型进行排名。

5 基线

本节将介绍一种用于在我们的基准数据集上进行评估的竞争性基线方法，称为关键实体跨媒体推理网络（KECMRN）。在介绍 KECMRN 之前，我们首先介绍其基本组件 KECMR 层。KECMR 层是一种模块化组合，由一个关键实体提取（KEE）层和多个跨媒体推理（CMR）层组成。KEE 层和 CMR 层由来自 [9] 的注意力单元和前馈单元构成。随后，我们使用 KECMR 模块与其他层组合成 KECMRN。最后，我们在 VTQA数据集上评估 KECMRN。

5.1 注意力和前馈-单元

如 [9]，所示，注意力单元与前馈单元相结合，具有强大的表征和学习能力。我们采用相同的设置来构建我们的单元。

多头缩放点积注意力单元。缩放点积注意力的输入包括维度为d_k的查询和键，以及维度为d_v的值。我们计算查询与所有键的点积，将每个点积除以$\sqrt{d_k}$，再应用softmax函数，得到值对应的注意力权重。给定一个查询$q \in \mathrm{R}^{1\times d}$、$n$ 个键值对（分别组合成键矩阵$K \in \mathrm{R}^{n\times d}$ 和值矩阵$V \in \mathrm{R}^{n\times d}$），注意力特征的计算如下：

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

多头注意力将输入分成多个部分，并在每个部分上执行单独的注意力计算。注意力特征如下：

$$MH(Q,K,V) = Concat(\mathrm{h}ead_1, \dots, \mathrm{h}ead_h)W^O \quad (2)$$

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (3)$$

其中，$W_i^Q, W_i^Q, W_i^Q \in \mathrm{R}^{d\times d_h}$ and $W^O \in \mathrm{R}^{h*d_h\times d}$ 是投影矩阵。

前馈单元。前馈单元接收多头注意力层的输出特征，并通过两个全连接层进一步进行变换

-层之间采用 ReLU 激活函数。

$$FFN(x) = \max(0, xW_1 + b_1)\,W_2 + b_2 \quad (4)$$

虽然在不同的 n_位置上，它们使用的参数因层而异。

5.2 关键实体提取层

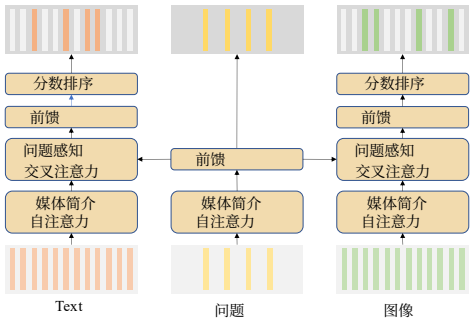

图 4: 关键实体提取层。

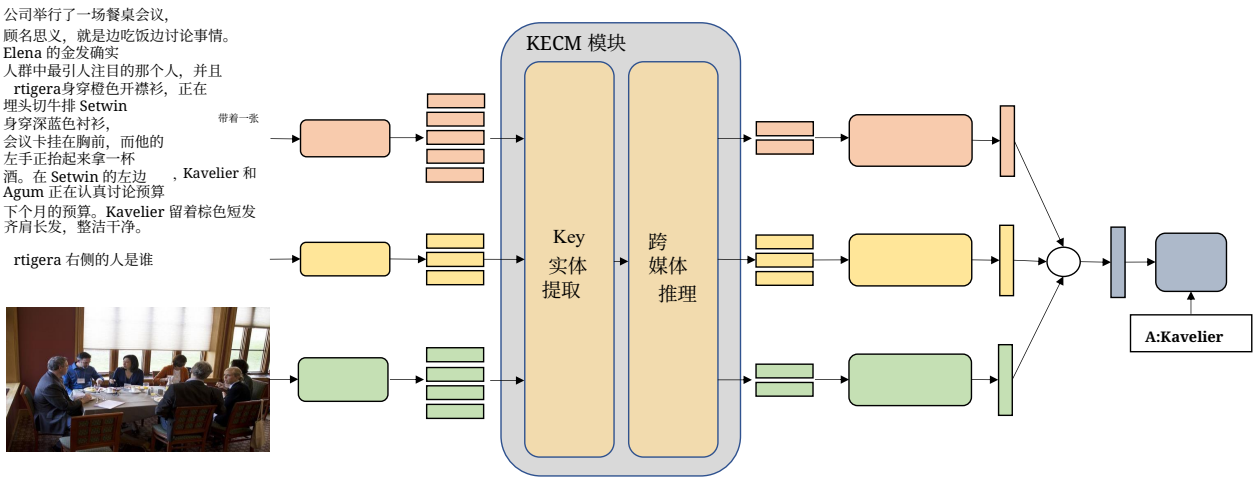

图 5： 关键实体跨-媒体推理网络的整体流程图。

如第 2 节所述，回答 VTQA 问题的第一步是根据问题找出关键实体。我们组合注意力单元和前馈单元，分别将问题信息整合到文本和图像中。然后，我们对问题感知文本/图像特征应用全连接层，为每个特征计算得分。最后，我们提取得分最高的 k 个特征作为关键实体。

完整的 KKE 层如图 4 所示。给定输入文本特征 $T \in \mathrm{R}^{l_t \times d}$、图像特征 $I \in \mathrm{R}^{l_i \times d}$ 和问题特征 $Q \in \mathrm{R}^{l_q \times d}$，KEE 层可表示为：

$$
\begin{aligned}
Q &= FFN\big(MH(Q,Q,Q)\big) \\
T &= FFN\big(MH(MH(T,T,T),Q,Q)\big) \\
I &= FFN\big(MH(MH(I,I,I),Q,Q)\big) \\
score_T &= W_T T + b_T \\
score_I &= W_I I + b_I
\end{aligned}
\tag{4}
$$

最终文本特征和 top-k 特征被视为图像/文本关键实体。为统一表达，我们将所有问题特征视为关键实体。

5.3 跨-媒体推理层

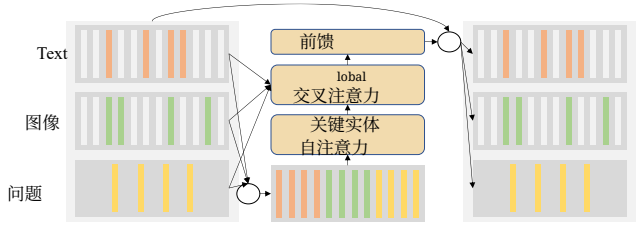

图 6： 跨-媒体推理层。"G"表示收集，"S"表示散布。

CMR 层旨在实现跨媒体的多步推理。如图 6 所示，我们首先从输入特征中收集关键实体，得到$S \in \mathrm{R}^{l_k \times d}$。然后，关键实体依次经过自注意力、基于原始特征的全局交叉注意力，以及前馈单元。最后，将关键实体分散到输入特征中。为满足多步推理的需求，我们在 KECMR 模块中插入了多个 CMR 层。CMR 层可表示为：

$$
\begin{aligned}
S &= gather([T,I,Q]) \\
S &= FFN\big(MH(MH(S,S,S),[T,I,Q],[T,I,Q])\big) \\
[T,I,Q] &= scatter(S,[T,I,Q])
\end{aligned}
\tag{4}
$$

5.4 关键实体跨-媒体推理网络

我们采用与 [10] 相同的网络框架，但用我们的 KECMR 模块替换 MCA 层，并添加一个额外的文本流。如图 5 所示，输入图像以自底向上方式 [11]表示为一组区域视觉特征。输入的问题和文本经过一个

-层 LSTM 网络 [12]。然后，我们多次使用 KECMR 模块来提取关键实体，并进行多步跨媒体推理。最后，我们使用注意力降维层融合多媒体特征，并将融合后的特征投影到答案概率分布中。

5.5 基线结果

我们在实验中使用的模型超参数如下。输入图像特征 d_i、输入问题特征 d_q、输入文本特征 d_t 和融合多模态特征 d_z 的维度分别为 2,048、512、512 和 1,024。多头注意力中的潜在维度 d 为 512，头数 h 设为 8，每个头的潜在维度 dh = d/h = 64。每个 KECMR 模块中的 CMR 层数以及 KECMR 模块的数量均为 2。

我们在训练集上训练模型 13 个轮次，并在开发测试集和测试集上对模型进行评估。结果如表 1 所示。

表 1： 我们的 KECMRN 在开发测试集-和测试集上的结果。

数据集 精确匹配 是非准确率 实体 F1 全局 F1				
开发测试集	0.527	0.782	0.619	0.527
Test	0.513	0.776	0.605	0.511

6 行政详情

我们已建立挑战赛网站[1]，并承诺至少在未来 3 年内维护该网站，同时持续更新数据集，以确保我们的多媒体问答挑战赛在该领域保持竞争力。我们将尽最大努力与 ACM多媒体大会的组织者合作，向研究人员宣传重大挑战赛任务，邀请他们参与。组织者和程序委员会成员名单如下，所有列入程序委员会的成员均已同意担任挑战赛专题投稿论文的审稿人。组织者。

• 吴向前。xqwu@hit.edu.cn。哈尔滨工业大学。• 陈康。

程序委员会。

• 车万翔。哈尔滨工业大学。• 陈康。哈尔滨工业大学。• 崔一鸣。哈

参考文献

[1] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D. 2017. 让 VQA 中的 V 发挥作用：提升视觉问答中图像理解的作用。载于 2017 IEEE计算机视觉与模式识别会议（CVP），fv6325-6334。[2] Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Hengel, A. 2017. Fvqa：基于事实的视觉问答。IEEE模式分析与机器智能汇刊，40(10)，2413-2427。[3] Marino, K., astegari, M., Farhadi, A., Mottaghi, . 2019. 知识型视觉问答：一项需要外部知识的视觉问答基准数据集。载于 IEEE/CVF计算机视觉与模式识别会议论文集，3195-3204。[4] Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H. 2017. 你比六年级学生更聪明吗？面向多模态机器阅读理解的六年级课本问题回答。载于 IEEE计算机视觉与模式识别会议论文集，4999-5007。[5] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K. W., Zhu, S. C., ... Kalyan, A. 2022. 学习解释：通过思维链进行多模态推理以回答科学问题。载于神经信息处理系统进展。[6] Talmor, A., Yoran, ., Catav, A., Lahav, D., Wang, Y., Asai, A., ... Berant, J. 2021. MultiModalQA：针对文本、表格和图像的复杂问题回答。载于国际学习表征会议。[7] Reddy, . ., ui, X., Li, M., Lin, X., Wen, H., Cho, J., ... Ji, H. 2022. MuMuQA：通过跨媒体知识提取与归纳进行多媒体多跳新闻问答。载于AAAI人工智能会议论文集，11200-11208。[8] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., amanan, D., ... Zitnick, C. L. 2014. Microsoft COCO：上下文中的常见物体。载于计算机视觉——ECCV 2014：第十三届欧洲会议，瑞士苏黎世，2014年9月6日至12日，论文集，第 V 部分 13，740-755。[9] Vaswani, A., Shazeer, N., Parmar, N., szkoreit, J., Jones, L., omez, A. N., ... Polosukhin, I. 2017. 注意力就是一切。神经信息处理系统进展，30。[10] Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q. 2019. 用于视觉问答的深度模块化协同注意力网络。载于 IEEE/CVF计算机视觉与模式识别会议论文集，6281-6290。[11] Anderson P, He X, Buehler C, et al. 2018. 用于图像描述和视觉问答的自下而上与自上而下注意力[C]//IEEE计算机视觉与模式识别会议论文集。6077-6086。[12] Hochreiter, S., Schmidhuber, J. 1997. 长短期记忆。神经计算，9(8)，1735-1780。

[1] https://visual-text-QA.github.io/